W198–W204 *Nucleic Acids Research, 2014, Vol. 42, Web Server issue* *Published online 30 May 2014*
doi: 10.1093/nar/gku427

R. S. WebTool, a web server for random sampling-based significance evaluation of pairwise distances

Florent Villiers[1,*], Olivier Bastien[2,*] and June M. Kwak[3]

[1]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20740, USA, [2]Laboratoire de Physiologie Cellulaire et Végétale, iRTSV, CEA-Grenoble, 17 rue des Martyrs, 38054 Grenoble Cedex 9, France and [3]Center for Plant Aging Research, Institute for Basic Science, Department of New Biology, DGIST, Daegu 711-873, Republic of Korea

Received January 30, 2014; Revised April 24, 2014; Accepted May 2, 2014

ABSTRACT

Pairwise comparison of data vectors represents a large part of computational biology, especially with the continuous increase in genome-wide approaches yielding more information from more biological samples simultaneously. Gene clustering for function prediction as well as analyses of signalling pathways and the time-dependent dynamics of a system are common biological approaches that often rely on large dataset comparison. Different metrics can be used to evaluate the similarity between entities to be compared, such as correlation coefficients and distances. While the latter offers a more flexible way of measuring potential biological relationships between datasets, the significance of any given distance is highly dependent on the dataset and cannot be easily determined. Monte Carlo methods are robust approaches for evaluating the significance of distance values by multiple random permutations of the dataset followed by distance calculation. We have developed R. S. WebTool (http://rswebtool.kwaklab.org), a user-friendly online server for random sampling-based evaluation of distance significances that features an array of visualization and analysis tools to help non-bioinformaticist users extract significant relationships from random noise in distance-based dataset analyses.

INTRODUCTION

To compare large data vectors, such as gene expression levels across multiple conditions, one needs to evaluate how similar they are. Various metrics are more or less commonly used that can be classified into two categories: correlation coefficients (e.g. Pearson, Spearman) and distances (e.g. Euclidean, Minkowski). As they do not yield identical results, they are intended to address different biological questions, and the appropriate metric to be used is highly dependent on both the working hypothesis and the original dataset (1,2). For instance, the Pearson correlation coefficient is the method of choice when dealing with absolute values (3), while Euclidean distance is described to perform better for the comparison of log2 ratio datasets (3). The fact that correlation coefficients are not sensitive to scaling is also a critical aspect when selecting a comparison method. Furthermore, distances allow for relative flexibility in the way the data are compared. Euclidean distance is more generally employed, but the use of the Minkowski distance with high power values, for example, has been mentioned to work better when analyzing high-dimensional datasets (4). This is due to the 'curse of dimensionality' (5), which describes the loss of contrast between entities of high dimension that is caused by a narrowing of the distribution of distances relative to their average (in other words, the variance remains constant while the mean increases proportionally to the square [for the Euclidean distance] of the dimension) (6). This phenomenon might be partially counteracted by increasing the power value of the Minkowski distance, which therefore appears to be a relevant approach for many biological, genome-scale datasets (4).

Distance values can be any positive real number, while the space of possible values for a correlation coefficient is constrained to the interval [−1,1]. This allows for the calculation of significance values corresponding to correlation coefficients (given a sample size N), which is difficult to achieve using a distance value. In addition, the validity of statistical inferences from parametric statistical tests (such as *t*- or *F*-tests) depends on various assumptions including normality, sufficient sample size and the fact that the experimental individuals have been chosen by taking random samples from well-defined populations. Neverthe-

*To whom correspondence should be addressed. Tel: +1 (301) 405 9727; Fax: +1 (301) 314 1248; Email: villiers@umd.edu
Correspondence may also be addressed to Olivier Bastien. Tel: +33 4 38 78 38 55; Fax: +33 4 38 78 50 91; Email: olivier.bastien@cea.fr

less, all three assumptions are rarely true in biological contexts. Samples from gene expression or metabolomics experiments, for instance, are usually acquired by non-random selection and then divided by randomization into experimental groups. Therefore, theoretical assumptions for parametric tests are rarely fulfilled, and it is theoretically invalid to use the classical *t*- or *F*-tests to analyze the experimental results.

With the continuous increase in computing power, stochastic approaches, known to quickly become computer-intensive, have emerged as robust and powerful ways of extracting meaningful information from random noise out of complex datasets (7), including distance matrices. Randomization tests belong to the broad class of Monte Carlo methods that rely on repeated random sampling to obtain numerical results. They are often used in computational biology (8) and physics (9) to address mathematical problems where it is difficult or impossible to obtain a closed-form expression or where a deterministic algorithm cannot be applied. The primary goal of randomization approaches is to test a null hypothesis (in this context, null is distinctly different from what it would be with a parametric test). The hypothesis is that there is a tendency for a certain type of pattern to appear in the data. The null hypothesis states that if this pattern is present, then it is purely a chance effect of observations in a random order. There is no requirement to have random samples from one or more populations. The following are the primary advantages afforded by the use of a stochastic method. (i) It does not assume that the data have any particular distribution. (ii) It applies to any test statistic. (iii) Unlike the sign and rank tests, it does not discard any information.

Here we present R. S. WebTool (http://rswebtool.kwaklab.org), a user-friendly platform for Monte Carlo-based significance evaluation of pairwise distances. The website implements a set of visualization tools and clustering algorithms and is intended to allow researchers with little to no experience in statistical methods to compute *P*-values associated with their pairwise distance of choice for a particular dataset. Publication-quality graphs are automatically generated, and all the output files can be retrieved for further analyses using third-party software. The R script is also fully available upon request for more advanced use and customization.

CONCEPTS AND METHODS

Server setup

While the overall structure of the website is written in PHP, the core code for distance calculation, random permutation, *P*-value computation, hierarchical clustering, adjacency function computation and static graph output is entirely written in *R* (10). Dynamic visualization tools are coded in JavaScript and are compatible with any browser without the need of external plugins.

Distance calculation

Distance matrices are generated using the built-in *dist* function of *R* (10), which are then vectorized. Euclidean distance $d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$ is arguably the most widely used measure of dissimilarity between vectors in biology and is therefore selected by default for one-click analyses. The Manhattan distance $d(p,q) = \sum_{i=1}^{n}|p_i - q_i|$, also called city-block distance, produces similar results to the Euclidean distance, although less sensitive to outliers. Both metrics are actually specific cases of the more general Minkowski distance $d(p,q) = \sqrt[k]{\sum_{i=1}^{n}|p_i - q_i|^k}, k = 1$. Although less commonly used in biology, additional metrics such as Canberra (11) and Chebyshev (12) distances are also available. Table 1 presents a summary of the available distances, together with examples of their use in the literature.

Random permutation and significance evaluation

Exact permutation tests are designed to make statistical inferences under the randomization model although the conclusions apply only to the results of experiments actually performed (13,14). By permuting the statistic of interest, such as the difference between means, distance, etc., the probability is calculated that the observed difference or a more extreme one could have occurred by chance (14,15). In this case, each data vector is independently shuffled by random sampling of the entire vector (all values are resampled) without replacement (each value can only be sampled once).

A statistic *S* is chosen to measure the extent to which the data show pattern A. The value *s* of *S* for the observed data is then compared with the distribution of *S* that is obtained by randomly reordering the data. A typical randomization test procedure includes (i) randomly shuffling the data many times, (ii) each time *i*, recording the value s_i^* of the test statistic *S* and (iii) computing the *P*-value by one of the following methods.

Evaluation of distance significances using discrete values. The *P*-value (or significance level of *s*) for a randomization test is the proportion of the shuffled test statistics that are more extreme in absolute value than the observed statistic (15). We will call this estimation the raw *P*-value. In mathematical terms, this is the estimation of $(S \geq s)$, the probability that *S* (test statistic, a random variable) is larger than or equal to the observed statistic *s*. Most of the time, the cumulative distribution function $F(s)$, where $F(s) = P(S \leq s)$, is unknown, and *P*-values have to be obtained by randomization tests.

Density-based P-value computation. Direct estimation of the raw *P*-value possesses one major drawback. If the real (unknown) *P*-value is too small, as in a biological sequence randomization test (16) where the *P*-value can often range from 0.1 to the order of magnitude of 10^{-30}, the number of randomizations can be too small to estimate the raw *P*-value (17). A classical method to avoid this drawback is to first directly estimate a density distribution function $\rho(s) = \frac{dF}{dS}(s)$ from the randomized statistic s_i^* by a kernel density estimation (18,19) and then to estimate the *P*-value from numerical integration of $\rho(s)$ (20,21) using the *R* built-in *nls* function.

Bandwidth optimization. A key feature of R. S. WebTool is the optional optimization of the density bandwidth for

Table 1. Summary of the distances that can be used with R. S. WebTool

Distance	Equation	Example of use	References
Manhattan (Minkowski, $P = 1$)	$d(p,q) = \sum_{i=1}^{n} \lvert p_i - q_i \rvert$		
Euclidean (Minkowski, $P = 2$)	$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$	Although the Euclidean distance remains largely preferred in gene expression dataset analysis, the use of higher power values is becoming more frequent as a way to overcome the 'curse of dimensionality'. On the other hand, lower power values (Manhattan distance or fractional values) allow for less sensitivity to outliers, and should be preferred in cases where they might interfere with the results.	(4,5)
Minkowski	$d(p,q) = \sqrt[k]{\sum_{i=1}^{n} \lvert p_i - q_i \rvert^k}$		
Canberra	$d(p,q) = \sum_{i=1}^{n} \frac{\lvert p_i - q_i \rvert}{\lvert p_i \rvert + \lvert q_i \rvert}$	The Canberra distance has been successfully used as a measure of stability for biological indicators identified from microarray data.	(11)
Chebyshev (Minkowski, $P = \infty$)	$d(p,q) = \max_i(\lvert p_i - q_i \rvert)$	Seldom used in biology, the Chebyshev distance can be utilized in gene set analysis and gene selection, in particular when UPGMA approaches are conducted.	(12)

normalized datasets. On-the-fly computation of the appropriate density bandwidth for each distribution using Scott's rule of thumb (22) works effectively in most cases ('Nrd' bandwidth setting from the analysis options panel). However, as it relies in part on the standard deviation of the distribution, very narrowly distributed distances after randomization (Figure 1A) appear to yield extremely low bandwidth, resulting in an 'over-reactive' density kernel estimator (Figure 1B) and therefore incorrect calculated *P*-values. To address this issue and prevent misinterpretation of the results, the *R* script implements a method for detecting such behavior and optimizing the density bandwidth (Figure 1C) by iterative adjustment until the standard deviation of the density falls below 1 ('Optimized' bandwidth setting from the analysis options panel).

Clustering

A very common step in a gene expression analysis workflow is the clustering of various genes or experimental conditions that are compared for function prediction (23) or identification of signalling cross-talk (21). While this server does not implement all possible algorithms for the numerous clustering methods, the script includes a hierarchical clustering step using *P*-value-based single, average and complete linkage followed by tree cutting at the following heights: $h = 0.05,\ 0.01,\ 10^{-3},\ 10^{-4},\ 10^{-5},\ 10^{-6}$. The resulting clusters can be mapped to the output weighted matrix and network view by entity reordering and color-coding, respectively.

P-value–distance function

The adjacency function that maps the interval of distances $[0,\infty]$ into $[0,1]$ can be computed if the user has selected data normalization at the time the job is launched. The dataset then becomes usable in analysis frameworks for weighted networks assuming pairwise edges of values constrained to the interval $[0,1]$ (24). The relationship *P*-value versus distance is therefore inferred by a logistic regression. Let $R(x,y)$ be a possible relationship between two vectors x and y (genes, biological samples, etc.). Then we can have $R(x, y) = 1$ if x and y are in relation (respectively, $R(x, y) = 0$ if x and y are not in relation). Let $D(x, y)$ be a predictive variable (for instance, a distance computed between x and y). We have the *a priori* probability $P(R(x, y) = 1)$ (respectively, $P(R(x, y) = 0)$)) to take the value 1 (resp. 0) and the conditional probabilities $P(D(x, y) \leq d | R(x, y) = 1)$ and $P(D(x, y) \leq d | R(x, y) = 0)$ of distance d conditionally to the $R(x,y)$ values. The logistic regression is based on the hypothesis that we have the relation $\log\left(\frac{P(D(x,y)\leq d|R(x,y)=1)}{P(D(x,y)\leq d|R(x,y)=0)}\right) = a + bd$ for some a and b. The general sigmoid function $p = A\frac{1}{1+e^{\frac{B-d}{C}}}$ is used for non-linear model fitting of *P*-value–distance with $A = 1$, where P is the *P*-value and d is the distance.

WEBSITE USAGE

Performing a new analysis is a three-step process that includes (i) dataset upload, (ii) options setting and (iii) analysis run.

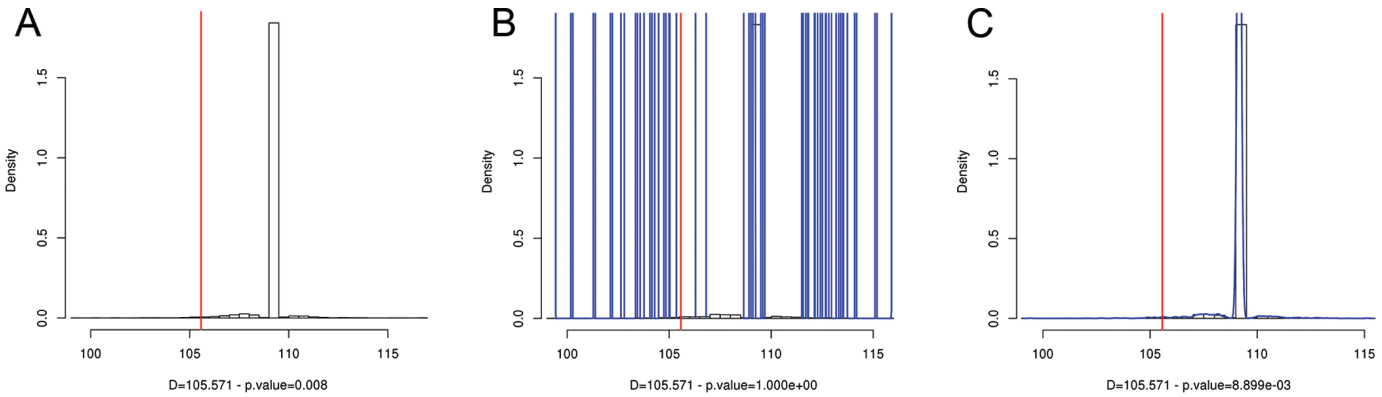

Figure 1. Density bandwidth optimization for narrowly distributed distances. The *P*-value is calculated from the original 'true' distance between the vectors (red) compared to the distribution of distances after random permutation of the dataset (black histogram). When computed, the density curve is displayed in blue. (**A**) Raw (non-density-based) computation yields a *P*-value of 0.008. (**B**) Density estimation using the 'Nrd' bandwidth from Scott's rule of thumb. Using this method gives a bandwidth of 2.85e-14, resulting in an over-reactive density kernel estimator. The computed *P*-value from density integration is 1. (**C**) Density estimation using the optimized Nrd bandwidth, which iteratively adjusts the density kernel bandwidth until the standard deviation of the density falls below 1. The new bandwidth is 0.109, yielding a *P*-value of 8.899e-3. The presented data are from the hormone dataset example file (normalized dataset, ABA 30 min versus ACC 1 h).

Inputs

The input dataset is in tab-delimited format and can be either copied and pasted into a text field or directly uploaded from a file. The columns represent the different data vectors to be compared with the first row considered the column names. Empty cells are allowed, provided they do not account for more than 50% of a data vector; the first column may contain characters other than numbers, in which case it will be assumed to be the row names and ignored when running the job. A typical file could contain expression values for a set of genes (rows) in various biological samples (columns). Such a file can be seen from the example dataset, which the user can fetch and use as a demo dataset. Once uploaded, the dataset is checked for errors before the user is directed to the next step (analysis options settings).

Although the default settings (Euclidean distance, no density calculation) allow for immediate submission of the new job, the user is also given the option of tweaking the way his dataset will be handled. Distances that can be used include the most common ones.

Outputs and visualization tools

Pairwise P-*value graphs.* The first panel focuses on pairwise significance evaluation of the different data vector combinations. Two types of R-generated, publication-quality graphs are displayed here, and dynamic filtering of the graphs (leftmost section) allows for quick retrieval of a particular comparison. Boxplot panels (Figure 2A, upper carousel) provide a per-vector overview of its similarity to each of the other vectors. The boxplots represent the distribution of the computed distances (after random permutation of the dataset) with which the original distance, displayed as a red star, can be compared. The reference data vector for a particular boxplot panel is specified on top of the graph. Using this view, the user may immediately identify the various data vectors that are similar to a particular one and evaluate to what extent this similarity is significant. For a more accurate view of each dyadic comparison, the lower graph carousel presents the distribution of the distances obtained after random sampling for each pair of vectors (Figure 2A). It is displayed as a histogram along with the original pairwise distance (red line) and, if computed, the density curve (in blue) of the distance distribution. Both the original distance and its associated *P*-value are displayed at the bottom of the graph. All images are available for individual download from this first panel, but can also be obtained as a batch from the download panel.

Integrated visualization tools

The subsequent panels present the dataset from a more integrated standpoint. The adjacency matrix (Figure 2B) displays the *P*-values of all the combinations in a color-coded matrix. Red squares indicate *P*-values that are lower than a user-specified threshold (significantly similar vectors given the specified metric), while blue squares indicate significantly different vectors over the same threshold (*P*-value $>$ 1 – threshold). Entering 0.5 as a threshold will enable all the squares to be displayed and can therefore be used to avoid hard filtering of the displayed dataset. All squares can be clicked to show the corresponding vector pair and its associated *P*-value. By default, the vector names are displayed in alphabetical order, and the user may reorder the dataset to match any of the 18 different clustering methods that have been pre-computed.

The network view (Figure 2C) is complementary to the matrix view. It displays each data vector as a node, while pairwise similarities are represented by the edges. Only those edges matching user-specified criteria are displayed, and pre-computed clusters are mapped to the network by node color-coding. The SIF (Simple Interaction Format) file associated with the displayed network is available for use in more advanced network visualization software such as cytoscape or gephi.

Dendrogram visualization of the dataset is also available (Figure 2D), taking the computed *P*-values or the normalized distances, if available, as input. The corresponding R-

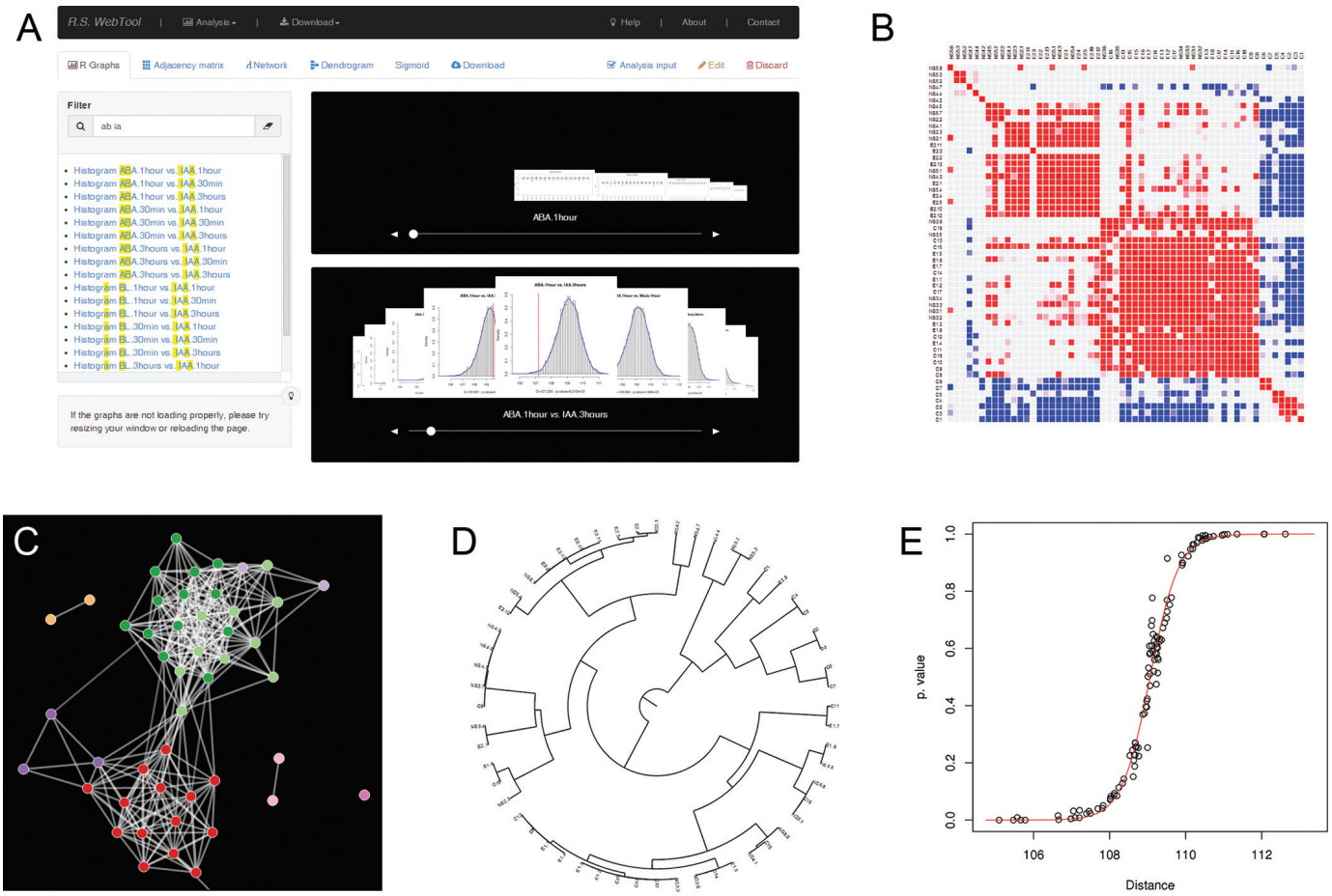

Figure 2. Overview of the various outputs and visualization tools available from R. S. WebTool. (**A**) The first panel focuses on pairwise comparisons: the upper carousel features boxplot panels that provide a one-to-all overview of the distance value significance for each entity (each column of the dataset versus all others). The lower carousel presents one-to-one comparisons for each pair of entities. In both types of graphs, true (original) distances are displayed in red, while distributions of the distances obtained after random permutation of the dataset are in black. (**B**) Adjacency matrix of the pairwise *P*-values resulting from the Monte Carlo analysis. Significantly similar entities (over a user-provided threshold) are in red, and significantly dissimilar entities are in blue. Dynamic cluster mapping can be achieved by reordering the rows and columns to highlight groups of similar entities. (**C**) Weighted network representation of pairwise distance value significance. Pre-computed clusters are mapped by color-coding the nodes (entities), and dyadic significance values above a user-provided threshold are represented by edge thickness. (**D**) Dendrogram representation of the adjacency matrix. This graph highlights potential clusters as well as the degree of similarity between them. (**E**) *P*-value–distance plot. If requested at the time the analysis has been launched, the adjacency function (red curve) for mapping the space of possible distances $[0,\infty]$ to the interval $[0,1]$ of *P*-values is computed from the plot. The equation of the deduced non-linear model and the coefficient of determination R^2 are provided. The presented screenshots are from various datasets, including those from the demo pages of R. S. WebTool.

generated Newick statement is shown along with the tree for use in third-party software. As a common way of displaying entity relationships, trees can be particularly meaningful in terms of visual clustering of the various elements to be compared as well as evaluation of inter-cluster relationships.

Downloadable files

All the generated files and datasets are made available for download. They include in particular the normalized input dataset (if normalization was requested), the adjacency matrix and the distances file (both original distances and those after random permutation). These files are in tab-delimited format. Newick and SIF files can also be retrieved for more advanced tree construction and network analysis, respectively, and R-output graphs (histograms, boxplot panels and adjacency graph) can be obtained as a batch from the download panel as well.

Personal information and data retrieval

Each analysis is given a unique ID containing a set of random characters that is used to generate a permalink. All analyses will remain on the server for at least a year and can therefore be accessed using this link. Although not required, users are encouraged to provide a valid email address to easily retrieve their previous work. To this end, a link toward a summary of all the performed analyses corresponding to a particular email address is sent upon user request.

DISCUSSION

We have developed the R. S. WebTool to address the specific issue of evaluating the significance of a distance value. Al-

though it is more targeted toward biologists, the proposed platform can handle any type of data, and its visualization tools are generic enough to be useful to a broad range of scientists. The main objectives of this tool are to provide researchers with (i) the possibility of applying a hard threshold for filtering significantly similar data vectors (using a standard *P*-value of 0.05 or 0.01, for instance) using a similarity metric of their choice other than a correlation coefficient and (ii) an adjacency function for mapping the domain of possible distances $[0,\infty]$ into the interval [0,1] when using soft thresholding approaches (or other types of analyses), assuming constrained pairwise similarity values are to be used (24). In the latter case, additional adjustment functions can be employed to ensure compliance of the dataset with the requirements of subsequent analytical methods, such as correlation coefficient-like values by using the function $x = 1 - 2y$ that maps the interval [0,1] into the interval [1,−1].

There are also some limitations to the use of R. S. WebTool. This, for instance, is the case when gene clusters are constitutively co-expressed, or when genes exhibit constitutive, nearly identical expression changes in response to numerous stimuli, but which relationship (e.g. inter-genes and inter-samples relationships, respectively) is not relevant to the biological question underlying the analysis. If these undesired dependencies account for a significant part of the dataset, Random Sampling-based analyses might produce low *P*-values that are not necessarily representative of a meaningful, biological interaction between the compared vectors with respect to the original question. For these very specific scenarios, one could simplify the dataset via Principal Components Analysis or any multivariate and multidimensional data analysis method, and use the generated eigenvectors (or eigengenes (25)) in subsequent Monte Carlo analyses.

Many gene function prediction algorithms heavily rely on expression profile similarity, which is almost always computed using a correlation coefficient (7,23). The possibility of evaluating the significance of a similarity value given by any distance metric, some examples of which have been shown to perform better than correlation coefficients in multiple cases (1–4), will increase the flexibility of the comparison methods as well as the accuracy of the prediction while still allowing one to make use of the available clustering/gene function prediction software that requires correlation coefficient-like data. The simplicity of the interface will make it easy for non-bioinformaticists to assess the significance of distance values from their dataset, while more advanced users can further process their results after downloading the generated files.

ACKNOWLEDGEMENT

The authors are grateful to Jade Lee and Roxane Bouten for critical reading of the manuscript.

FUNDING

National Science Foundation [IOS1025837 to J.M.K.]; Agence National de la Recherche [ANR-10-BLAN-1524 to O.B., ANR-12-BIME-0005 to O.B., ANR-12-JCJC to O.B.]; Institut National de la Recherche Agronomique [FUGAL to O.B.]. Funding for open access charge: Institute for Basic Science, Republic of Korea.
Conflict of interest statement. None declared.

REFERENCES

1. Ramoni,M.F., Sebastiani,P. and Kohane,I.S. (2002) Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 9121–9126.
2. Troyanskaya,O., Cantor,M., Sherlock,G., Brown,P., Hastie,T., Tibshirani,R., Botstein,D. and Altman,R.B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.
3. D'Haeseleer,P. (2005) How does gene expression clustering work? *Nat. Biotechnol.*, **23**, 1499–1501.
4. Gibbons,F.D. and Roth,F.P. (2002) Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.*, **12**, 1574–1581.
5. Aggarwal,C.C., Hinneburg,A. and Keim,D.A. (2001) *Database Theory — ICDT 2001*. Vol. 1973, In: Bussche, J. and Vianu, V. (eds), Springer, Berlin, Heidelberg, pp. 420–434.
6. Lespinats,S. (2006) Style du génome exploré par analyse textuelle de l'ADN. *Ph.D. Thesis*, University Paris 6.
7. Luo,F., Yang,Y., Zhong,J., Gao,H., Khan,L., Thompson,D.K. and Zhou,J. (2007) Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics*, **8**, 299.
8. Bastien,O., Aude,J.C., Roy,S. and Marechal,E. (2004) Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics. *Bioinformatics*, **20**, 534–537.
9. Macgillivray,H.T. and Dodd,R.J. (1984) Monte-Carlo simulations of galaxy systems. *Astrophys. Space Sci.*, **105**, 331–337.
10. Team,R.D.C. (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
11. Jurman,G., Merler,S., Barla,A., Paoli,S., Galea,A. and Furlanello,C. (2008) Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics*, **24**, 258–264.
12. Nowakowska-Zajdel,E., Mazurek,U., Stachowicz,M., Niedworok,E., Fatyga,E. and Muc-Wierzgon,M. (2011) Cellular signal transduction pathways by leptin in colorectal cancer tissue: preliminary results. *ISRN Endocrinol.*, **2011**, 575397.
13. Ludbrook,J. (1994) Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. *Clin. Exp. Pharmacol. Physiol.*, **21**, 673–686.
14. Ludbrook,J. and Dudley,H. (1998) Why permutation tests are superior to *t* and *F* tests in biomedical research. *Am. Stat.*, **52**, 127–132.
15. Manly,B.F.J. (2006) *Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition*. 3rd edn. Taylor & Francis, London, England .
16. Ortet,P. and Bastien,O. (2010) Where does the alignment score distribution shape come from? *Evol. Bioinformatics Online*, **6**, 159–187.
17. Moore,D.S. and McCabe,G.P. (2006) *Introduction to the Practice of Statistics*. W.H. Freeman & Company, New York, NY.
18. Parzen,E. (1962) On estimation of a probability density function and mode. *Ann. Math. Statist.*, **33**, 1065–1076.
19. Rosenblatt,M. (1956) Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27**, 832–837.
20. Press,W.H. (2007) *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY.
21. Villiers,F., Jourdain,A., Bastien,O., Leonhardt,N., Fujioka,S., Tichtincky,G., Parcy,F., Bourguignon,J. and Hugouvieux,V. (2012) Evidence for functional interaction between brassinosteroids and cadmium response in Arabidopsis thaliana. *J. Exp. Bot.*, **63**, 1185–1200.
22. Scott,D.W. (2009) *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York, NY.
23. van Noort,V., Snel,B. and Huynen,M.A. (2003) Predicting gene function by conserved co-expression. *Trends Genet.*, **19**, 238–242.

24. Zhang,B. and Horvath,S. (2005) A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article17.

25. Langfelder,P. and Horvath,S. (2007) Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.*, **1**, 54.